Rule-Based and Statistical Morph Segments in English-to-Finnish SMT [*]

Tommi A Pirinen
tommi.pirinen@computing.dcu.ie
ADAPT Centre—School of Computing, DCU

Antonio Toral
atoral@computing.dcu.ie
ADAPT Centre—School of Computing, DCU

Raphael Rubino
rrubino@prompsit.com
Prompsit Language Engineering S.L.

February 2, 2016

Abstract

Morphological segmentation is recognised as a potential solution in statistical machine translation (SMT) to deal with data sparsity posed by morphologically complex languages like all Uralic languages. Two approaches have been used in the literature, rule-based and statistical, but always in isolation. In addition, previous work has failed to bring significant improvement and conclusive analyses of the effects of segmentation. In this paper we use both rule-based and unsupervised approaches to segmentation jointly and aim to find out where they excel and where they fail. Our case study is on English-to-Finnish using the datasets provided at the WMT 2015 shared task. We present a comprehensive evaluation of SMT systems built with different segmenters including: intrinsic evaluation, MT automatic metrics, MT human evaluation and MT linguistic evaluation. In terms of automatic metrics, the best system is the one that combines both rule-based and unsupervised segmentations, outperforming an unsegmented system by 1.08 BLEU and 3.64 TER points. Human evaluation shows that the outputs produced by an SMT system with rule-based segmentations are preferred over those of the system that uses unsupervised segmentations.

1 Introduction

Morphologically complex languages are well known to cause problems for contemporary *statistical machine translation* (SMT) systems. *Morphological segmentation*, in which words are divided into sub-word units prior to training, has been a popular method to deal with morphologically complex languages in SMT. In this regard, [1] presents a comprehensive overview of the topic. However, despite a lot of effort put

[*] This article has a free/open publication licence; this particular version is Tommi A Pirinen's draft version. This version uses my own documentclass instead of official publication's, if any. This version is optimised for screen reading. This version uses only free fonts if possible. This work is licensed under a Creative Commons Attribution–NoDerivatives 4.0 International Licence. Licence details: `http://creativecommons.org/licenses/by-nd/4.0/`. Original publication in proceedings of second IWCLUL held in Szeged 2016

into the use of morphological segmentation in SMT, automatic evaluation for morphologically rich languages, to the extent of the richness of Uralic languages rather than most Indo-European, have yielded modest improvements to downright negative results [2].

In this paper we aim to lay out a pathway on this problem by extensively studying various segmentation schemes and the errors they introduce and avoid. We systematically evaluate and compare segmentations produced by the two most widely used approaches to morphological segmentation: rule-based and statistical. Our aim is then to find out where each of these approaches excels and where they fail, and whether their joint use can be beneficial. To have a systematic and thorough evaluation, we use four evaluation schemes for the segmentations: intrinsic test of segmentation quality against a gold standard, the automatic and human evaluation of segmented MT systems, automatic evaluation of linguistic features and translation model features.

2 Morphological Segmentation

Morphological segmentation is a well-established technique in SMT and there is a large amount of related work to consider: [3, p. 324] provides an extensive reading list on the topic. It is important to note that the term *morphological segmentation* is often used for a number of different techniques, which we believe are not comparable, or relevant to Uralic SMT. For example, segmenting Chinese non-space separated sentences or Vietnamese syllables into words is not considered in this article (we associate this to *tokenisation*), nor changing word-forms into structure consisting of abstract dictionary word and suffix identifiers (i.e., *morphological analysis*). The segmentation we consider involves solely finding segmentation points within a word token. There are two approaches to segmentation that we study in this article: unsupervised statistical and rule-based. The state-of-the-art of statistical segmentation has been determined in *MorphoChallenge* shared tasks[4]. In rule-based morphology, researchers generally concentrate on the higher level linguistic morphological analysis rather than on plain segmentation, and thus there is no comprehensive evaluation of the state of the art in the *segmentation* task. For Uralic languages it makes sense to follow the prevalent *Finite State Morphology* [5]. These are the frameworks we use in this paper.

Much of the prior work in Finno-Ugric languages, mainly for Finnish and Estonian, is based on using unsupervised morphological segmentation only [2, 6, 7, 8, 9]. One of the new emphases presented in this article is the comparison and combination of rule-based morphologies to unsupervised segmentation. [7] make use of rule-based segmentation in determining their baseline but they carry on to use only the better-performing unsupervised segmentation in their actual experiments.

The majority of prior work that shows more optimistic scores concerns language families whose morphological complexity is considerably simpler than that of Uralic languages, e.g. Slavic [10] and Germanic [11]. German is mainly pre-processed for compound simplification; Finnish in comparison is productive both for compounding and for regular inflection. The closest language that has been extensively studied in previous work is Turkish [12, 13]. In addition, Basque [14] is comparable to Finnish in terms of morph distributions after segmentation.

Our approach to morphological segmentations is done in pre-processing and post-processing steps that perform addition and deletion of segmentation points operating on segmentation markers. We compare two approaches to morphological segmentation: rule-based and unsupervised. The methods are implemented using the following

Segmenter	text
None	kuntaliitoksen selvittämisessä
hfst-comp	‘kunta→←liitoksen selvittämisessä
hfst-morph	kunta→←liitokse→←n selvittämise→←ssä
Flatcat	kun→←tali→←itoksen selvittämis→←essä
Morfessor	kun→←ta→←liito→←ksen selvittä→←misessä
Gloss	municipality+annexation.Gen examination.Ine
Translation	examination regarding municipal annexation

Table 1: Different segmentation methods.

software: HFST [15][1] for rule-based and Morfessor [16] for unsupervised segmentation. For both approaches we create two different segmentation models: for rule-based we select segmentation points to match annotations for word-segmentation (referred to in the rest of the article with the code-name `hfst-comp`) and morph-segmentation (`hfst-morph`). For unsupervised versions we use Morfessor 2.0 Baseline (`morfessor`) and Morfessor Flatcat (`flatcat`) [17]. Our experiments include morphological segmentation methods used separately as well as in system combination.

For rule-based morphological segmentation we developed a segmenter on top of omorfi [18],[2], an open-source implementation of weighted finite-state morphology. Omorfi's morphological segmenter has a number of segments annotated: `MB` for morph boundaries, `DB` for derivation boundaries, `WB` and `wB` for word boundaries and `STUB` for other stemmer-type boundaries. We use these to produce two segmented versions: one where all `WBs` and `wBs` are turned into segmentation points and one where `WBs`, `wBs` and `MBs` are.[3] These are called *compound* and *morph* segmentations, respectively. Rule-based morphological segmentation is ambiguous and we use the 1-best result. For word-forms not recognised by the morphological analyser, no segmentation points are produced.

Unsupervised morphological segmentation is based on statistically likely segmentation points found from an unannotated training corpus when trying to iteratively optimise a given function. In the case of `morfessor`, the optimisation function is based on minimum description length (MDL), so the aim of the algorithm is to minimise the vocabulary size of the output, i.e. to find the segmentation with the lowest number of different morphs. `Flatcat` extends this by using hidden markov models (HMM) and context to create classes for the morphs: stems, suffixes, prefixes and non-morphemes. Thus, for example, if there is a morph identified as a common suffix, it should be more unlikely for it to be split off from the beginning of a word, even if doing so would result in a lower set of distinct morphs as per MDL. For each word we select the 1-best segmentation.

Examples of the different segmenters are shown in Table 1. The semantics of the gloss can be most easily traced to match the `hfst-morph` version.

[1] `http://hfst.sf.net`

[2] `http://github.com/flammie/omorfi/`

[3] The two remaining types of segmentation points (`DB` and `STUB`) are discarded as they are not relevant for our task.

3 Experimental Setup

3.1 MT Tools and Datasets

Our experimental setup matches the one used in [19].The training, development and test data set used in our experiments are obtained from the WMT 2015 shared task.[4] In this shared task, participants train and apply MT systems on pre-defined corpora for training, development and testing, the domain of the latter being news. Our translation models (TMs) are trained on the Europarl v8 Finnish–English parallel corpus and the language models (LMs) benefit from the additional shuffled news monolingual corpus (*News Crawl: articles from 2014*). All typical pre-processing steps are performed. All the scripts used to pre-process the data are available with the Moses distribution [20]. Finally, we generate segmented training sets for both parallel and monolingual corpora following the segmentation methods described in Section 2. The segmented SMT systems output segmented Finnish text, thus a post-processing step (morph-joining) is performed to obtain the final translation.

We assess empirically the performance of two LM training methods: concatenation of parallel and monolingual corpora, or linear interpolation of two individual LMs based on the minimisation of the perplexity obtained on the development set. We observe that segmented LMs reach better results with the concatenation method, while the word-based LM benefits from the interpolation approach. We also experiment enriching the phrase-based SMT pipeline with additional components such as multiple reordering models (joint use of word-, phrase-based [21] and hierarchical [22] re-ordering models), Operation Sequence Model (OSM) [23] and neural language models [24] such as the Bilingual Neural LM (BiNLM) [25]. Again, we empirically evaluate adding these models to our SMT systems based on the development set. We observe an improvement of the results with the three reordering models for segmented and non-segmented systems, while OSM and BiNLM yield improvements to the word-based system only.

3.2 Segmentation

For our rule-based segmentation we used the segmentation automaton `omorfi.segment.hfst` from omorfi version 20150326 by simply rewriting the word boundary and morph boundary markers into arrows and any other boundaries into zero-length strings as described in Section 2. For unsupervised segmentation we used `morfessor` 2.0.2-alpha and `Flatcat` 1.0.4 trained on the Finnish side of the europarl-v8 corpus, using default settings except for the fact that we remove the segmentation points of non-morphemes in `Flatcat`.

Accordingly, we build four SMT systems using the aforementioned segmentations: `hfst-morph`, `hfst-comp`, `morfessor` and `Flatcat`. In addition, we explore the joint use of more than one segmentation by means of system combination with `MEMT` [26].[5] We try three different combinations:

a) `combo-unsup`, where we attempt to build the most competitive system using only unsupervised methods. This combines `morfessor`, `Flatcat` and the baseline SMT system (unsegmented). b) `combo-rb`, where the attempt is on building the most competitive system using rule-based methods. This combines `hfst-morph`, `hfst-comp`

[4] `http://www.statmt.org/wmt15/translation-task.html`

[5] We use default settings except for the radius (5, default is 7), following empirical results obtained on the devset.

System	F-Measure	Precision	Recall
Flatcat	54.04 %	76.04 %	41.91 %
hfst-comp	43.82 %	97.63 %	28.25 %
hfst-morph	86.32 %	92.39 %	81.00 %
Morfessor	53.89 %	71.01 %	43.42 %

Table 2: Results of the intrinsic evaluation of the four segmentation methods

and, again, the baseline. c) `combo-all`, where the aim is to build the most competitive system using both unsupervised and rule-based methods. This combines all the five systems: `morfessor`, `Flatcat`, `hfst-morph`, `hfst-comp` and the unsegmented baseline system.

3.3 Evaluation

For intrinsic evaluation of segmentation accuracy we used morphochallenge 05 [27] gold test data and `evaluation.perl`,[6] since it most closely resembles the segmentation setup of our SMT setting (i.e. no annotations or deep analysis).

Automatic evaluation of MT outputs was performed using the following evaluations scripts: `mteval13a.pl` for BLEU [28], `tercom-7.25.jar` for TER [29][7] and `meteor-1.5.jar` for METEOR [30].[8]

For human evaluation we used the Appraise toolkit [31].[9] The evaluation was conducted by three Finnish native speakers with a background in Computational Linguistics. This evaluation is inspired by the human evaluation conducted as part of the translation task in WMT; the evaluators are given a set of outputs coming from different systems and they are asked to rank them according to their quality (ties are allowed).

Finally, for the linguistic analysis, we used the morphological fluency classifications of [7], basing on those, we developed a new automatic evaluation script using omorfi analyses.

4 Evaluation

4.1 Segmentation Evaluation

In order to evaluate how the quality of segmentation –as defined by the gold standard written by a linguist– affects the final MT results, we evaluated our segmentation methods on the gold standard provided at the morphochallenge 2005 shared task on morphological segmentation. The results are shown in Table 2.

As expected, the results of the linguistic analyser `hfst-morph` matches the linguistic gold standard quite well, with unsupervised methods performing considerably worse. The linguistic analyser `hfst-comp` of course does not obtain high recall in segmenting all boundaries as it only aims to select a very specific subset of those (i.e. compound boundaries or stem-stem boundaries in unsupervised terms).

[6] `http://research.ics.aalto.fi/events/morphochallenge2005/data/evaluation.perl`

[7] `https://www.cs.umd.edu/~snover/tercom/tercom-0.7.25.tgz`

[8] `https://www.cs.cmu.edu/~alavie/METEOR/download/meteor-1.5.tar.gz`

[9] `http://github.com/cfedermann/Appraise`, commit 9b643ae55647...

	Dev. set			Test set		
System	**BLEU**	**TER**	**METEOR**	**BLEU**	**TER**	**METEOR**
Baseline	0.1577	0.7479	0.3069	0.1402	0.7609	0.2997
Flatcat	0.1481	0.7699	0.3060	0.1387	0.7712	0.3001
hfst-comp	0.1541	0.7415	0.3019	‡0.1471	0.7405	0.2977
hfst-morph	0.1575	0.7381	0.3050	‡0.1451	0.7476	0.2986
Morfessor	0.1434	0.7868	0.2987	0.1343	0.7882	0.2942
combo-unsup	0.1595	0.7267	0.3031	0.1408	0.7367	0.2937
combo-rb	0.1569	0.7179	0.3002	‡0.1459	**0.7214**	0.2959
combo-all	‡**0.1638**	**0.7160**	**0.3074**	‡**0.1510**	0.7245	**0.3011**

Table 3: Automatic evaluation of MT systems built with different segmentation methods. The baseline is unsegmented. Statistical significance tests (paired boostrap resampling) run on BLEU (‡ $p = 0.01$).

4.2 MT Automatic Evaluation

We evaluate MT systems built on the training data segmented using each of the four segmentation methods with the three aforementioned state-of-the-art automatic metrics: BLEU, TER and METEOR (see Table 3).

We observe that systematically the system combination of all segmentation models performs the best, with the exception of TER on the test set, where the combination of rule-based and baseline methods results in the best score. Furthermore we note that the rule-based combination beats the unsupervised combination on the test set, but on the dev. set the unsupervised combination is slightly better (except for TER). Contrasting this to single system scores, which are worse across the board, we can conclude that each individual system contributes different parts to the output produced by the system combinations.

4.3 MT Human Evaluation

We performed human evaluation of the translations with 3 native speakers ranking the sentences. We produced the final rankings from the human evaluation judgements using the TrueSkill method adapted to MT evaluation [32] with its implementation in WMT-Trueskill,[10] following its usage at WMT15.[11] Namely, we run 1,000 iterations of rankings followed by clustering ($p = 0.95$). Results are shown in Table 4.

#	Score	Range	System
1	0.529	1-2	combo-all
2	0.414	1-2	combo-rb
3	-0.943	3-3	combo-unsup

Table 4: The results of human evaluation by three native speakers with background in computational linguistics as measured by TrueSkill.

The results show that human annotators, in general, prefer either the combination of all systems (`combo-all`) or the rule-based combination (`combo-rb`) over the purely unsupervised combination (`combo-unsup`). More specifically, `combo-all` is the best performing system (0.529), closely followed by `combo-rb` (0.414) with `combo-unsup` clearly performing worst (-0.943). In terms of significance (column range), at $p = 0.95$, `combo-all` and `combo-rb` are in the same cluster (range 1-2), thus meaning neither of the two is significantly better than the other, while `combo-unsup` is in a different cluster

[10] `https://github.com/keisks/wmt-trueskill`

[11] `https://github.com/mjpost/wmt15`

System	**NM**	**TP**	**POSS**	**NAA**	**SVA**	**PP**
Frequency	10.03	1.48	1.40	0.92	0.76	0.14
Baseline	71.94	**39.24**	54.83	31.62	45.22	21.97
Unsup	72.38	37.24	**60.36**	33.80	**46.78**	21.94
Rule-based	72.95	36.32	56.65	32.42	45.13	**29.49**
Combo	**73.34**	36.78	56.06	**34.37**	43.87	24.26

Table 5: Linguistic fluency of translated sentences compared to the reference translation. The metric is F_1 of the analysed MT output compared to the analysed reference. Frequency is the number of occurrences of the construction (as automatically detected) in the reference translation per sentence.

(range 3-3), meaning its performance is significantly worse, compared to the other two systems.

The inter-annotator agreement as shown by Fleiss' $\kappa = 0.26$ suggests that there is a mild tendency of agreement between the annotators. This is in the same range as agreement at the WMT 2014 shared task [33].

4.4 MT Linguistic Evaluation

In order to evaluate the fluency of the translations, [7] suggest using morphological analysis to determine translation issues over a set of linguistic criteria. We measure the recall of the following constructions in the MT output as compared to the reference translation: a) *Noun marking* (NM), for nouns with case different than nominative. b) *Possession* (POSS) for any word with possessive suffix. c) *Noun-adjective agreement* (NAA) for sequences of adjective-noun, where case is shared. d) *Subject-verb agreement* (SVA) for sequences of noun-verb, where number is shared. e) *Transitive object* (TP) for sequences of verb-noun, where case is accusative or partitive. f) *Postposition* (PP) for sequences of adposition-noun, where case is genitive. Of these tests, NM and POSS pertain to single tokens and NAA and SVA sequences of two tokens, whereas TP and PP scan the whole context and are thus less reliable.

There is no clear tendency for any single system to be the best in morpho-syntactic fluency as measured by these tests, e.g., it seems that combo and rule-based systems will recover NM and PP better but unsupervised matches the most POSS forms. An additional error analysis should reveal the effects of missing forms.

The translation models (phrase and reordering tables) present different characteristics whether the training data was segmented or not, but also according to the different segmentation methods. For instance, depending on the segmenter, the number of extracted and scored phrase-pairs in the phrase table differs, as shown in the first row of Table 6. These results show that segmenting the data leads to a larger amount of phrase-pairs extracted, which is related to the differences in alignment points found by MGiza. Only hfst-morph leads to a lower amount of extracted phrase-pairs. The performance of each segmentation method according to Table 2 is apparently inversely correlated with the number of phrase-pairs: the highest the *f-score*, the lower the amount of phrase-pairs.

An interesting phenomenon is observed on the word-level fertility from English to Finnish (how many Finnish words are generated by one English word), as shown in the second row of Table 6. These scores indicate to which extent the segmentation leads to ambiguous alignments. These results are supported by the lexical ambiguity scores shown in the third row of the same Table 6. The lexical ambiguity scores are obtained by averaging the number of target words aligned with a source word with a non-null

	Baseline	Flatcat	hfst-comp	hfst-morph	Morfessor
#Phrase-pairs (M)	84.6	86.2	86.8	82.6	85.5
Fertility	0.786	1.029	0.856	1.151	1.047
Lexical ambiguity	43.3	29.3	36.7	24.0	28.7

Table 6: Statistics extracted from the trained SMT models, the first row indicates the number of phrase-pairs (millions), the second row contains the word-level fertility measured (English→Finnish) and the third row indicates the average number of target words aligned with each source word calculated at the corpus level.

Source	The water should be conducted to a fixed drain or rain water network, and not just into a container.
Baseline	Vesi pitäisi johtaa kiinteään viemäriin tai että sadevesi verkkoon, eikä vain astian.
hfst-morph	Vesi pitäisi hoitaa kiinteään viemäriin tai **sadevesiverkostoon**, eikä vain astiaan.
Reference	Vesi pitäisi johtaa kiinteään viemäriin tai sadevesiverkostoon, eikä vain astiaan.
Source	the news is reported by BBC, who refers to governmental sources.
Baseline	uutinen on raportoinut BBC, joka viittaa valtion lähteistä.
hfst-comp	uutinen on raportoinut BBC, joka viittaa **hallituslähteisiin**.
Reference	asiasta kertoo BBC hallituslähteisiin viitaten.

Table 7: Examples of translations where words in bold are generated at decoding time without being observed in the training data.

probability at the corpus level, the lower the score the better. We can see that the fertility scores are inversely correlated with the lexical ambiguity. These notable differences between our SMT systems lead to variable translations from the same source sentences depending on the SMT system used. To illustrate these differences, we show some translation examples in Table 7. As shown in the examples, the morph-based translation methods can come up with a correct compound or morphological combination, not found in training data, e.g., the term *sadevesiverkostoon* (sewage network) rather than the un-idiomatic and grammatically questionable *sadevesi verkkoon* (network of rainwater). In the second example the generated compound *hallituslähteisiin* matches the idiomatic compound for ‘governmental sources’ whereas the baseline results in the less idiomatic *valtion lähteistä* ‘sources from the state’ and gets the case wrong.

5 Conclusions and Future Work

This paper has explored the joint use of different segmentations methods in SMT for the English-to-Finnish language direction. We have shown that both rule-based and unsupervised morphological segmentation methods are useful as they are complementary. While morphological segmentation approaches in isolation do not result in substantial increments of performance according to automatic MT metrics, using different segmentations jointly does lead to notable increments of performance (+1.08 BLEU and -3.64 TER compared to an unsegmented system).

For future work it might be interesting to see if some of the more advanced morphological processing methods. For example abstraction of morphemes and morph prediction method used by [7] has been shown to improve English-Finnish translation. Likewise, using n-best lists and re-ranking with morphs—e.g. in style of [34, 8]—could improve the final system even more.

Regarding the automatic system that uses a morphological analyser to check for linguistic similarity, for future research it would be interesting to couple this with a syntactic parsing in order to better recognise long-span features such as verb argument structures.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

We thank Inari Listenmaa, Hege Roivainen and the student organisiation of linguistics in University of Helsinki for human evaluation.

References

[1] Ann Clifton. *Unsupervised morphological segmentation for statistical machine translation*. PhD thesis, Applied Science: School of Computing Science, 2010.

[2] Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*, pages 491–498, Copenhagen, Denmark, September 2007.

[3] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[4] Mikko Kurimo, Mathias Creutz, and Ville Turunen. Overview of morpho challenge in clef 2007. In *Working Notes of the CLEF 2007 Workshop*, pages 19–21, 2007.

[5] Kenneth R Beesley and Lauri Karttunen. *Finite State Morphology*. CSLI publications, 2003.

[6] Mark Fishel and Harri Kirik. Linguistically motivated unsupervised segmentation for machine translation. In *LREC*, 2010.

[7] Ann Clifton and Anoop Sarkar. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 32–42, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 148–157, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[9] Adrià De Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73–76, 2009.

[10] Maja Popovic and Hermann Ney. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of LREC*, 2004.

[11] Fabienne Cap, Alexander Fraser, Marion Weller, and Aoife Cahill. How to produce unseen teddy bears: Improved morphological processing of compounds in smt. In *Proceedings of EACL 2014*, 2014.

[12] Kemal Oflazer and Ilknur Durgar El-Kahlout. Exploring different representational units in English-to-Turkish statsistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, 2007.

[13] Coşkun Mermer. Unsupervised search for the optimal segmentation for statistical machine translation. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pages 31–36, 2010.

[14] Relevance of different segmentation options on Spanish-Basque SMT, author=de Ilarraza, Arantza Dıaz and Labaka, Gorka and Sarasola, Kepa, booktitle=Proceedings of the EAMT, year=2009.

[15] Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A Pirinen, and Miikka Silfverberg. Hfst—framework for compiling and applying morphologies. *Systems and Frameworks for Computational Morphology*, pages 67–85, 2011.

[16] Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. Morfessor 2.0: Python implementation and extensions for morfessor baseline. 2013.

[17] Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1177–1185, 2014.

[18] Tommi A Pirinen. Omorfi–Free and open source morphological lexical database for Finnish. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, page 313, 2015.

[19] Raphael Rubino, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. AbuMaTran at WMT 2015 translation task: Morphological segmentation and web crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 184–191, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[20] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, 2007.

[21] Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *IWSLT*, pages 68–75, 2005.

[22] Michel Galley and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, 2008.

[23] Nadir Durrani, Helmut Schmid, and Alexander Fraser. A joint sequence translation model with integrated reordering. In *Proceedings of ACL/HLT*, pages 1045–1054, 2011.

[24] Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. Decoding with large-scale neural language models improves translation. In *EMNLP*, pages 1387–1392, 2013.

[25] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1370–1380, 2014.

[26] Kenneth Heafield and Alon Lavie. Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme. *The Prague Bulletin of Mathematical Linguistics*, 93:27–36, 2010.

[27] Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy, and Murat Saraçlar. Unsupervised segmentation of words into morphemes–challenge 2005: An introduction and evaluation report. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, 2006.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318, 2002.

[29] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for machine translation in the Americas*, pages 223–231, 2006.

[30] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[31] Christian Federmann. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September 2012.

[32] Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[33] Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014.

[34] Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. Technical report, DTIC Document, 2008.